

Name of the session	Towards the implementation of High Value Datasets – First hands on
Organizer	Roberto Di Bernardo (Engineering Ingegneria Informatica SpA)
Description of the session & speaker(s)	
<p>The Implementing Act on High Value Datasets from the European Commission mandates public sector bodies to publish open, machine-readable data belonging to specific thematic categories (including statistics, Earth Observation, mobility). The stress is firmly put on facilitating the reuse of such datasets even (but not only) for economic purpose; the Act lays down the arrangements for publishing and reusing high-value datasets, the applicable conditions for re-use and the minimum requirements for disseminating data via APIs; aiming also at using open datasets through AI-based applications, it is also important to define and make available the terms of use of the APIs and the related quality of service.</p> <p>The session explored how projects funded under the topic Public Sector Open Data for AI (DIGITAL-2022-CLOUD-AI-02-OPEN-AI) are doing their first steps to support the actual implementation High Value Datasets in EU territories for the benefit of local communities.</p> <p>The initial introduction by the moderator, Roberto Di Bernardo, gave the general context of the session and explained the overall rational of the agenda; it presents a wider perspective putting the Public Sector Open Data for AI projects in relation with the more complex picture that includes Public Sector Open Data CEF initiatives and the Digital Europe Programme ones, thinking of Data Spaces particularly.</p> <p>Then two projects where presented:</p> <ul style="list-style-type: none"> - BeOpen: “How to support local communities with HVDs- initial steps from different EU countries” presented by Francesco Mureddu (The Lisbon Council) - RODEO: “Open Data platform for exchange of meteorological HVDs” presented by Minna Huuskonen <p>After the presentation of BeOpen and RODEO, the moderator also gave few details about the other two sister projects currently running: Maregraph and OME2. The four projects are dialoguing about a cluster collaboration.</p> <p>The panel session followed. The discussion, as anticipated, enlarged the scope by also involving representatives from:</p> <ul style="list-style-type: none"> - SPOTTED project (CEF-TC-2020-2 Public Open Data) – Antonio Filograna (Engineering Ingegneria Informatica SpA) - GREAT project (DIGITAL-2021-CLOUD-AI-01-PREP-DS-GREEN-DEAL) – Mark Dietrich (EGI Foundation) - DS4SSCC project (DIGITAL-2021-CLOUD-AI-01-PREP-SMART-COMM) – Martin Traunmueller (Austrian Institute of Technologies) <p>At the end, final remarks, conclusions and thanksgivings to the speakers, panellists, and audience were formulated by the moderator.</p>	

Main topics of discussion

The first presentation focused on RODEO as an open data platform for the exchange of meteorological High-Value Datasets (HVDs). It briefly introduced the Open Data Directive (EU) 2019/1024 and the Implementing Act on High Value Datasets. The speaker emphasized the significance of the latter in particular concerning technical requirements for the categories of HVDs, it envisages, to adhere the FAIR principles, ensuring that the data is usable, accessible, interoperable, and reusable.

RODEO focuses on Meteorological HVDs.

Here the implementing regulation provides technical parameters to be respected (e.g. frequency and timeliness). The presentation showed the key objectives of the project aligned with EUMENET strategy and the World Meteorological Organization's requirements (WIS2.0). To increase the open availability of meteorological data, with the expected level of quality, across the Europe to boost their usage within new digital products for better warnings, forecasts, and services to the public and weather-critical industries and to contribute to the safe and efficient functioning of society with multiple benefits across the European economy, industry, and society.

A presentation of BeOpen project followed. General information was provided, highlighting in this case a focus on a wider spectrum of HVD categories to support the 8 different Smart Cities and Territories use cases going to be implemented by the project in 6 different EU countries. Key areas are: urban mobility and street management; natural disaster management; urban security and infrastructure management; invading species detection/protection. Technically, the project starts for the results of several CEF Public Open Data projects (such as: Spotted; INTERSTAT; GreenMov) with a strong focus on context data interoperability (FIWARE based). The challenges are to make a convergency among different nature of Open Data particularly concerning the updated frequency, moving from static information towards near real-time one (e.g. those extracted from Urban Data Platform) and to assure the right level of quality and reliability. Strong attention is going to be paid on the one side on the legal constrains that can prevent public authority open data publication and from the other side on legal obligation the published HVDs have to meet. Thinking of AI-based services

After RODEO and BeOpen presentation, the moderator briefly introduced the other two sister projects: Maregraph (focusing on marine related datasets) and OME2 (following the activities of Open Maps for Europe, focusing on EuroGeographics datasets).

The second part of the session started with the panel.

The first question was about SPOTTED project and the innovation it is trying to carry on concerning HVDs. Key goal is facilitating PAs with the publication of open data that can support and improve the management of green areas in cities. Different sources of information are considered, particularly Earth Observation (Copernicus) services. Data are used to evaluate a green index that measures the quality of green areas.

Some possible synergies were identified in the debate:

- with OME2-SPOTTED: administrative boundaries among city areas are not easy to be clearly identified.
- with RODEO-BeOpen: meteorological HVDs play an important role in climate change related use cases addressed by BeOpen

For data reusability, key aspects are accessibility, but also data reliability and quality. The latter may sensibly differ from one source to another, so it should be mandatory to perform a quality check before a dataset is published.

The discussion then moved on the gaps in HVDs identified by the directive and those that markets and stakeholders require.

From GREAT project (preparatory action for green deal data spaces), having to objective to create a framework for green deal datasets, the stress is put on the consistency of HVDs among the different Member States in Europe and to push towards the use of common (standard) APIs. It is also a matter of alignment of local, national, and global effort: a multi-scale alignment is a key factor in achieving the sharing of quality and consistent data. Some help in this sense can be provided by a close collaboration among the different projects working on the topic. Clearly, projects like RODEO and OME2 can play an important role due to the participation of Central Public Authorities that can support the uptake of common recommendations collected by the ongoing projects towards the uptake of common approaches. Even more important if we consider the full picture also composed by the TEFs and Local Digital Twins (currently the Commission is publishing tenders aiming at the releasing of a Local Digital Twin Toolbox).

In line with this, the fifth project involved in the discussion, DS4SSCC (preparatory action for data spaces), whose focus is on the development of a blueprint for sustainable and smart city/community, including its governance, technologies, and relevant datasets. For priority datasets, the followed methodology starts from the scanning of all the use cases related to smart cities and communities from the different countries involved and prioritize them (e.g. geographic spread, connection, interoperability mechanisms, data quality), and then proceed with the selection of the most relevant ones to elicit similarities and datasets. Recommendations for HDVs should come out together with guidelines for replication.

During the panel, feedback and questions from the audience were continuously collected and discussed. Among the risen points:

- Even though we are talking about Public Sector Open Data, there should be some mechanism in place to limit the use of datasets for specific security issues (e.g. hostile countries), when it comes for potential critical data (e.g. land use). In some case, there are already some approaches that offer different level of data granularity both in terms of time and spatial resolution, but for sure this is a topic that requires a further investigation.
- Datasets quality check is an important aspect. How to balance the need for more datasets available (e.g. increasing the use of crowdsourced ones) with the related cost required? From RODEO experience, at the moment, the main quality control is performed over metadata and there are clear standard for its exchange.
- Linked with the previous point, it is important to consider the overall sustainability, so specific business models should be defined and applied. Freemium approach can be considered offering different levels in terms of granularity, resolution, frequency.
- The projects should provide some reference data management pipelines to support stakeholder replication activities. It was mentioned INTERSTAT project (CEF Public Open Data) that provides methodology and pipeline for statistical open data management following standards and consolidated approaches. Here again the importance of involving central public authorities (in this case ISTAT, INSEE, and EUROSTAT) to potentially boost replicability not only at National level, but also across Member States.

Main outcomes

- There is the need to assure homogeneity and consistency of datasets and related metadata among the different stakeholders at all levels within a single Member State and in between Member States, particularly in terms of standards used and quality.
- There is the need not only of tools to manage and publish HVDs, but also of agreed and potentially widely reusable pipelines for HDVs management, in particular to guarantee the adequate level of quality.

- To assure the reusability of HVDs, AI related challenges, like overfitting, must be considered and addressed.
- The session was the first action of a future collaboration between the Public Open Data for AI projects. The collaboration is envisaged to share approaches, methods, and datasets to boost EU HVDs consistency and reusability among the different stakeholders in EU Member States.
- In a wider picture, the session also highlighted importance of a collaboration between the Public Sector Open Data for AI projects and preparatory actions for data spaces (SSCC and Green Deal) and the future data spaces that will be deployed. A two-way exchange of roadmaps and guidelines should be put in place before the end of the preparatory actions.

Contribution of the session to the Strategic Agenda/Roadmap for Data and AI

- Boost multi stakeholder collaboration around new data regulations – the session showed a strong alignment of intents towards the actual implementation of HVDs by the different projects involved, going beyond the cluster of Public Sector Open Data for AI projects, but also considering data space preparatory actions. This will help in eliciting both common good practices and possible gaps still to be covered.
- support the increasing of data business opportunity understanding – the session more than once put the accent on the sustainability of the overall HVDs management pipeline: to find the right balance between assuring the right quality of datasets (and metadata) and keeping the overall process affordable, in particular for local public administrations.
- Achieve global impact and maximize societal & Environmental benefits – AI and Data technologies are recognised as a valid support to boost the EU GDP and to achieve climate change mitigation. The session put the basis for a multifaceted and multi-level collaboration between different projects focusing on: better events forecast and management; more sustainable local ecosystems; land monitoring and planning.
- Support the increase of societal trust in AI and Data – the technology is not fully accepted by the society (at different levels) and unfounded mistrust can limit its uptake. The session showed the strong commitment of all the projects involved to release HVDs with the appropriate level of quality and reliability, keeping in mind that trustful AI solutions can be obtained only starting from well governed and trusted data management processes.